AUTOMATING CONVOY TRAINING ASSESSMENT TO IMPROVE SOLDIER PERFORMANCE

Noelle LaVoie*
Parallel Consulting
Longmont, CO, 80501

Peter Foltz, Mark Rosenstein, Rob Oberbreckling
Pearson Knowledge Technologies
Boulder, CO 80301

Ralph Chatham
ARPA Consulting
Falls Church, VA 22043

Joe Psotka
U.S. Army Research Institute
Arlington, VA 22202-3926

ABSTRACT

Monitoring teams of decision-makers in complex military environments requires effective tracking of individual Soldier and team performance. An untapped source of timely and diagnostic performance information lies in ongoing communications among Soldiers operating as a team. With the right analyses the communication data can be connected to both the team's and each individual's performance, abilities and knowledge. The DARCAAT program developed and tested a toolset for automating team assessment and near real-time alarms. The toolset uses Automated Speech Recognition and Statistical Natural Language-based techniques for embedding automatic, continuous, and cumulative analysis of team communication in training and operational environments. Based on the toolset, applications were developed that apply the metrics and models to support After Action Reviews (AARs) and real-time alarms.

1. INTRODUCTION

There are numerous challenges to effectively identify, track, analyze, assess, and report on team performance in near real-time in complex training and operational environments. For example, current methods of assessing team performance often rely on temporally delayed outcomes or global metrics. These metrics often lack information rich enough to diagnose failures, detect critical incidents, or provide feedback on needed improvements. However, the content of the information communicated by teams provides detailed indicators of the information team members know, what they tell others, and their current situation. Using this information, it is possible to derive powerful indicators of team performance based on real-time data available in communication.

1.1 Automated Communication Analysis

Verbal communication provides a rich source of information about a team's performance, including what team members know, how information flows through the team's network, and detailed information about cognitive states, situation awareness, workload and stress. In fact, within the team training community, trainers and subject matter experts often rely on listening to a team's communication to assess how well a team is performing. In order to exploit the information inherent in verbal communication, technologies are needed that can assess the content and patterns of the verbal information flowing in the network and then convert this information into variables that can support straightforward, usable feedback for teams and commanders as well as alarms to indicate when a team may be heading into trouble.

The overall goal of automated verbal communication analysis is to apply a set of computational modeling approaches to networked communication to convert the verbal communication into useful characterizations of performance. These characterizations include metrics of team performance, feedback to commanders, and alerts about critical incidents related to performance. This type of analysis has several prerequisites. The first is the availability of sources of clear verbal communication. Second, performance measures which can associate the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **DEC 2008**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**
4. TITLE AND SUBTITLE **Automating Convoy Training Assessment To Improve Soldier Performance**		5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Parallel Consulting Longmont, CO, 80501**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release, distribution unlimited**		
13. SUPPLEMENTARY NOTES **See also ADM002187. Proceedings of the Army Science Conference (26th) Held in Orlando, Florida on 1-4 December 2008, The original document contains color images.**		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	

communication to actual team performance are needed. Finally, these prerequisites can be combined with computational approaches applied to the communication in order to perform the analysis. These computational approaches include computational linguistics methods to analyze communication, machine-learning techniques to associate communication to performance measures, and finally cognitive and task modeling techniques.

1.2 Communication Analysis Pipeline

By applying this combination of computational approaches to team communication, we have developed a complete communication analysis pipeline (see Figure 1). Communications are converted directly into performance metrics which can then be incorporated into visualization tools to provide commanders and Soldiers with applications such as automatically augmented AARs and debriefings.

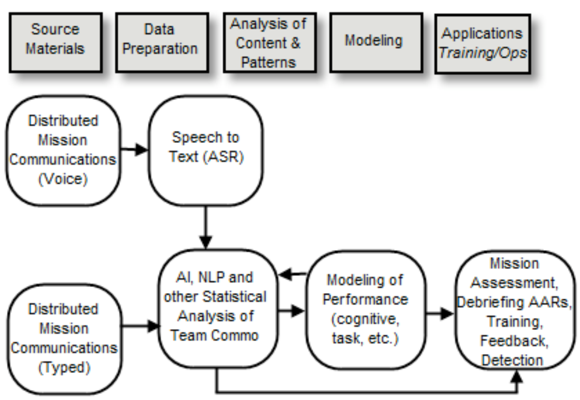

Figure 1. The communication analysis pipeline.

Individual components of the communication analysis pipeline have been previously researched and tested. Over a series of studies, computational language-based communications methods have been evaluated favorably in terms of their ability to predict team performance. For instance, they are successfully able to predict team performance scores in simulated task environments based only on communications transcripts (Foltz, Lavoie, Rosenstein, & Oberbreckling, 2007; Foltz, 2005; Foltz, Martin, Abdelali, Rosenstein & Oberbreckling, 2006; Gorman, Foltz, Kiekel, Martin & Cooke, 2003; Kiekel, Cooke, Foltz, Gorman & Martin, 2002; Kiekel, Gorman & Cooke, 2004). Using human and ASR transcripts of team missions in a UAV environment, in simulators of F-16 missions, and in Navy TADMUS exercises, the methods predicted both objective team performance scores and SME ratings of performance at very high levels of reliability.

The language analysis techniques have also previously been tested for the analysis of Automated Speech Recognition (ASR) input for a limited portion of a dataset of verbal communication (see Laham, Bennett & Derr, 2002 and Foltz, Laham & Derr, 2003). The results indicated that even with typical ASR systems degrading word recognition by 40%, the system's prediction performance degraded less than 10%. Thus, even with high ASR error rates, which are typical in live recordings, such a system can provide robust performance predictions.

2. DATA COLLECTION

Two datasets were collected and analyzed during this effort. In collaboration with the Fort Lewis Mission Support Training Facility, we collected audio data from the DARWARS Ambush! virtual environment convoy training. In Ambush! up to 50 Soldiers jointly practice battle drills and leadership during simulated convoy operations. At the National Training Center (NTC), Fort Irwin, a second dataset was collected consisting of data from live mounted convoy STX lane training. In collaboration with the NTC Observer/Controllers (O/Cs) performance assessments of the datasets and recorded AARs and hot washes from the live training exercises were collected. Both data collection efforts concentrated on platoon and squad-level teams performing convoy operations.

Both in Ambush! and at NTC units are trained in situations currently encountered on a daily basis by units deployed for Operations Enduring Freedom and Iraqi Freedom. In the training, the convoy commander conducts troop-leading procedures, issues a movement order, and leads the convoy along the designated route. The convoy encounters contacts along the route, which can include a civil disturbance, a rocket-propelled grenade attack, an improvised explosive device (IED), a near ambush, vehicle-borne IED (VBIED), negotiation with Iraqi police and complex attacks (IED and ambush) (see Kuhn, 2004).

2.1 DARWARS Ambush!

DARWARS Ambush! is a widely used game-based training system that has been integrated into training for many brigades prior to deployment in Iraq (Diller, Roberts, Blankenship & Nielson, 2004; Diller, Roberts & Wilmuth, 2005). In this environment up to 50 Soldiers are able to jointly practice battle drills and leadership training during simulated convoy operations. Figure 2 shows a typical user's view during training. At Fort Lewis, we were able to coordinate the collection of over 250 training.

Figure 2. DARWARS Ambush! training scenario screen.

2.2 National Training Center

Data collection at the NTC was significantly more challenging than collection of the Ambush! data, as might be expected from trying to instrument real platoons and squads in the field. We collected voice activated recordings of SINCGARS FM communications during STX lane training, although voice quality was not as high as in the controlled Ambush! environment.

Data was collected during rotations from January through June of 2007. We recorded a total of 105 STX lane training missions, of which we selected 57 recordings that had acceptable quality audio, and training events of interest. These recordings varied in duration from as little as ten minutes to several hours. Combined with the 250 missions recorded from Ambush! at Fort Lewis, we collected a total of over 300 training missions.

3. DEVELOPMENT OF PERFORMANCE METRICS

Providing feedback on team performance requires the toolset to automatically associate performance metrics with communication streams. Thus, the system typically requires one or more metrics of team performance, which can include objective measures of performance, such as threat eliminations or mission objectives completed, or subjective measures of performance, such as Subject Matter Experts' (SME) ratings of aspects of performance including command and control and situation awareness. In both the Ambush! and NTC convoy training contexts, evaluation occurred as part of the AAR process, so it was important that the performance measures were drawn from the same task context, and developed in conjunction with SMEs with extensive experience working with convoys.

We developed five scales that captured the important dimensions of performance in this domain based on a mission essential task list (METL) (see FM 3-0, Army Operations and FM 7-1, Battle Focused Training): command and control (C2), situation understanding (SA), adherence to standard operating procedures (SOP), critical action drills (CA) and general team performance (TEAM). The Army's standard three point rating scale of Trained, Practice, and Untrained was expanded into a five point scale anchored at the top (Trained), middle (Practice) and bottom (Untrained). Seven SMEs rated the audio collected from Fort Lewis and NTC on these scales, using a rating tool developed for the project that presented the audio in a visual interface, allowing SMEs to select segments of audio and complete their ratings. The SMEs were also asked to distinguish between critical events, defined as events that change the scope of battle, the commander's plan or disrupt the operational tempo, and other training events in the communication. Finally, SMEs conducted AARs for every mission they rated, providing sustains, improves and ratings for the entire mission.

Before using SME ratings as a performance measure, it is important to assess how well the SMEs agreed with each other. All SMEs were asked to rate a pair of missions selected for the purpose of collecting data to compute reliability and agreement. Intraclass correlations among the SMEs ranged from .76 to .85 ($p<.001$) for average items suggesting excellent reliability. The intraclass correlations for single items ranged from .38 to .66 ($p<.001$). Exact agreement (two SMEs agree on the exact score) was calculated between every pair of SMEs, and average exact agreement ranged from 24% to 50%. Average adjacent agreement (SMEs agree within one score point) ranged from 74% to 96%. Two SMEs had extremely high agreement, with their adjacent agreement ranging from 93% to 100%, and exact agreement ranging from 51% to 86%. The agreement among SMEs was impressive, and indicates that the SME ratings are appropriate for computational modeling. It also provides support that SMEs are able to accurately detect performance from communication.

4. MODELING APPROACH

In order to be able to process communication, technology is needed that can "understand" the meaning of what is being conveyed in the communication. The primary underlying technology used in this analysis is a method for mimicking human understanding of the meaning of natural language called Latent Semantic Analysis (LSA), (see Landauer, Foltz & Laham, 1998 for an overview of the technology, and Foltz, 2005 for its application to team

communication analysis). LSA is automatically trained on a body of text containing knowledge of a domain, for example a set of training manuals and/or domain relevant verbal communication. After such training, LSA is able to measure the degree of similarity of meaning between two communication utterances in a way that closely mimics human judgments. This capability can be used to understand the verbal interactions in much the same way that a subject matter expert can compare the performance of one team to others. The results from the LSA analysis are combined with other computational language technologies which include techniques to measure syntactic complexity, patterns of interaction and coherence among team members, audio features, and statistical features of individual and team language (see Jurafsky & Martin, 2000 for an overview of approaches to language analysis). These features include measures that examine how semantically similar a team transcript is to other transcripts of known quality, measures of the semantic coherence of one team member's utterance to the next, the overall cohesiveness of the dialogue, characterizations of the quantity and quality of information provided by team members, and measures of the types of words chosen by the team members.

The computational representation of the team language is then combined with machine-learning technology to predict team performance metrics. Machine learning techniques including hill-climbing methods such as stepwise regression, discriminant analysis, and Support Vector Machines (SVMs) are then used to determine the language features that best model the performance metrics without overfitting the data. Essentially, these methods learn which features of team communication are associated with the different performance metrics and then can predict scores for the team performance metrics for new sets of communication data.

5. ANALYSES AND MODELING RESULTS

To go from audio data and SME ratings to a system that can automatically rate new missions requires building predictive models of the data. The goals of modeling were to identify critical events in segments of audio communication and assess team performance to support automated AARs and identify critical events. Data modeling was conducted on a set of 72 training missions which included communication data, speech analysis variables, and SME-selected critical events and ratings of performance.

5.1 Automatic Speech Recognition

The automatic speech recognition (ASR) component was used to translate the audio into text and extract audio features. We used BBN Technologies' AVOKE STX speech-to-text software system. AVOKE transforms the raw, digitally recorded audio into a machine-readable text transcript for analysis. ASR systems, including AVOKE, require preliminary training in the domain of interest to produce reasonable recognition accuracy rates. The ASR system used here is trained on accurately human-transcribed audio recordings. The system may then inductively "learn" associations between features in the audio signal and the pre-transcribed words humans interpreted when they listened to and transcribed that signal. This process of learned association results in a trained language model which allows the ASR system to determine which words should be recognized from the audio features found in a sample of new audio.

In order to test the ASR performance, the system was trained on 16 hours of communication. A set of 802 utterances were held out from the ASR training set and this set was then run through the ASR system and compared against the human transcribed transcript. Word error rate, calculated as the sum of the insertions, deletions and substitution errors made by the ASR system divided by the total number of words, was 33.7%. This error rate is consistent with results found for the Speech In Noisy Environments (SPINE) evaluation (see Schmidt-Nielsen et al., 2001). Prior modeling work suggests that this range of error rate may decrease system prediction accuracy by only 10% from verbatim transcripts, which can still provide acceptable performance predictions (see Foltz, Laham & Derr, 2003).

5.2 Speech Feature Analysis

Voice stress analysis examines the physiological basis that changes in person's stress level causes micro-muscle tremors (MMT) in the vocal tract muscles. These MMTs can affect the energy and frequency of the speech signal, (see Lippold, 1971; Hanson et al., 2002). Voice Stress analysis has not been tested for predicting performance in teams, but seems likely to contain useful information for predicting performance. In team communication situations stress does not need to be hidden, and indeed may help to convey urgency, failures, or degree of criticality in a situation. Thus, with appropriate analyses it may be possible to detect stress features in team communication, leading to predictions about how a team is performing.

We used a number of statistical transformations of the speech signal to determine how likely it is that

stress is present in a segment of communication. Using variables derived from the speech samples Hidden Markov Models (HMM) were used to categorize speech as excited or neutral. The primary features that were used in the models were measures of power, pitch, change over time, frequency components (FFT), rate, duration and frequency of speech. Overall, the excitement classification algorithm worked well, with 87% accuracy for female voices and 81% accuracy for male voices. Being able to detect excitement in an utterance does not fully determine whether there is a critical event, or whether a team is performing poorly or well on a particular team performance metric. However, these results suggest that the method can provide useful information that can be incorporated with the other variables described below to help detect critical events and help tune the performance models.

5.3 Team Performance Modeling

Team performance modeling was performed to predict the SME ratings of performance based on variables drawn from the text of the communications, such as semantic content, as well as variables drawn directly from the audio features of the communication, as described above. During the rating process, SMEs identified spans of times as "events" and then provided ratings on the metrics for that event. Typical events ranged from a minute to five minutes in duration. Using a dataset divided up into events, we developed automated prediction models in which we trained the system on the communication of 80% of the events randomly chosen and then tested predictions on the remaining 20% of the events. The best variables were selected to predict the team's performance on each of the five scales for each training event.

Table 1. shows the correlation between the model's predicted rating and the SMEs ratings of the events using a model that combined text and speech variables. The model's predictions were correlated with the SME ratings between .36 and .43, somewhat lower than the agreement between SMEs which ranged from .38 to .66 for single items. Nevertheless, they do show that the model can provide fairly accurate predictions of a team's performance at the event level.

Table 1. Correlation Between SME Ratings and Model Predictions

Metric	R	N	p value
CA	.37	572	<.001
CC	.41	838	<.001
SA	.41	833	<.001
SOP	.43	886	<.001
TEAM	.36	799	<.001

Team performance was also modeled for entire missions, instead of the separate training events in the missions, based on the ratings of the two SMEs with the highest agreement. Because the unit of analysis for this model was the entire mission, and the agreement results for the SMEs were reported using events as the level of analysis, additional agreement measures were calculated based on the team performance ratings for entire missions rated by both of the SMEs. The model's predictions correlated well with the SME ratings, with correlations ranging from .70 to .81 across the five scales, only slightly lower than the correlations between the two SMEs. Adjacent agreement between the SMEs and the model was also quite high, strongly supporting the use of the model in the toolset for assessing a team's performance.

5.4 Critical Event Modeling

A critical event is anything that changes the scope of battle, the commander's plan or disrupts the operational tempo. Such changes are important in training since teams and/or commanders may not notice the change or may not respond appropriately to the change, so it is important to be able to identify critical events to assess performance as well as be able to later play back the events that lead up to the critical event for AARs. Critical event modeling was conducted using a spectrum method utilizing discrete time windows where the size of the window, and step size between windows, were optimized to predict critical events from the communication data. A support vector machine then classified the data into categories with a high or low probability that a given time window includes a critical event. Using this approach, over 80% of the critical events were detected with an acceptably low false alarm rate (ROC area under the curve was 95.6%). This model allowed the toolset to accurately detect critical events during a mission for inclusion in an AAR. In addition, the sensitivity can be adjusted, so that more critical events could be detected, although with higher levels of false alarms which may be useful if a commander wanted to be alerted to any kind of team anomaly, or in cases where sensitivity could be reduced so that commanders are alerted only if the system is highly confident that a critical event is occurring.

6. AAR TOOL DEVELOPMENT

Convoy training conducted at Fort Lewis using Ambush! and during STX lanes at NTC relies on the After Action Review process to maximize the benefits of training. During a well run AAR, the O/C or commander reviews the unit's performance,

emphasizing areas where the unit would benefit from improvement as well as areas the unit should sustain at their current high level of performance.

The value of being able to provide a unit with recorded examples of their performance is unquestionable. After several hours of training, many team members may not be able to accurately recall a particular incident from earlier in training in sufficient detail to be able to learn well from their experiences. Currently, some video and audio from training events are collected at the NTC. However, the video and audio are seldom available to units for AARs or hot washes conducted in the field. NTC is in the process of installing the necessary infrastructure to provide live video and audio feeds to the O/Cs in the field, including laptops carried in the O/Cs' vehicles and plasma displays available in trailers distributed through the training area. These improvements will make it possible for O/Cs to use the recorded media of a unit's training to augment the AAR process. Within DARWARS Ambush! it is possible to record a unit's performance as they navigate the challenges in the virtual world, and then play the video back during an AAR. But two obstacles remain, even if all the multimedia is available. The first is the time required in finding events noted as training relevant during the mission by sifting through the video and audio recordings and making sure they cover the "teaching points" that illustrate a unit's weaknesses. With current O/C staffing shortages, the time that it takes to identify segments of video or audio of interest may overwhelm the benefits of using recorded performance for AARs. The second obstacle is that given the workload for understaffed O/Cs not all activity can be continuously monitored and critical events may be overlooked. By automatically analyzing the communications, this toolset extends the O/Cs reach.

The AAR tool we developed includes several functions to support O/Cs and commanders in preparing an AAR. As shown in Figure 3, O/Cs can view an entire training mission by events. This view provides a color-coded table of automatically selected events and critical events that are rated by the tool on the 5 scales: CC (Command and Control), SA (Situation Awareness), SOP (Standard Operating Procedures), CA (Critical Action Drills), and TP (overall Team Performance). The lowest scores are indicated by red, with the best scores shown in green, to help O/Cs spot events of interest. Clicking on the rating scale name (e.g. CC) sorts the events so the events with the best or worst performance on that scale will be visible at the top (see Figure 3), making it easy for an O/C to identify potential sustains and improves. Each event is linked to the audio recording, so clicking the event will play the associated audio files automatically. Clicking the event will also display brief, automatically derived comments for each event that explain the event and ratings (see above right in Figure 3). As shown in the lower half of Figure 3, the display also allows O/Cs to browse using a timeline interface, with the ability to get an overview of the whole mission and zoom in to locate audio from particular parts of the mission they want to listen to.

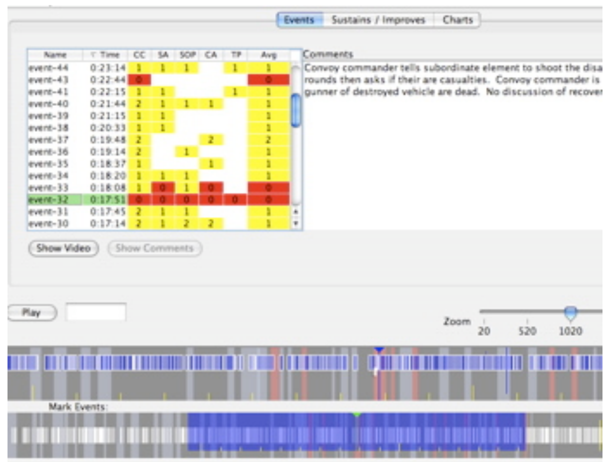

Figure 3. AAR tool interface showing events and ratings.

6.1 Evaluation of the AAR Tool

Two SMEs reviewed the AAR tool in order to provide us with feedback about its usefulness in supporting AARs, and to suggest improvements and other possible applications for the DARCAAT toolkit. The SMEs included our primary SME, LTC (Ret) Fena, and a second SME who had recently returned from his second tour as a convoy commander in Iraq. Both SMEs thought the AAR tool was valuable and would reduce the time required to prepare for an AAR as well as increase the scope of events that could be discussed. They emphasized that time is often the most precious commodity during training and the focus of the AAR tool should remain on shortening AAR prep time to maximize the tool's utility to O/Cs and commanders. Both SMEs thought that the tool layout was conducive to the way they would choose to use it. Specifically, they felt that the tool would allow a quick and easy three-step process for preparing an AAR:

1. Identify a unit's strengths and weakness at a glance, by scanning sorted event ratings;
2. Understand the weaknesses by examining these events in more detail, including listening to audio samples;

3. And last, pull all the information about the unit's performance together with their own comments.

The SMEs suggestions for improving the functionality of the AAR tool, included:

- Making critical events easier to find, either by creating a separate table for them or by marking them more clearly in the context of the other events;
- Allowing an O/C or commander to add their own comments to events and missions;
- Providing short descriptions of each event, such as "First IED" or "CASEVAC" to facilitate identification of events;
- Adding performance benchmarks to help standardize performance across units. They felt that rating a unit as "trained" on a particular metric, such as command and control, is often a subjective judgment, and the Army's training could benefit by calibrating the ratings provided by the AAR Tool to a more objective standard.

The SMEs also believed that the tool could easily be extended to provide an O/C or commander support beyond a typical training mission AAR. Their ideas for extending the tool included adding longitudinal tracking to monitor a unit's performance over multiple missions. This would require archiving missions and adding tools to visualize and summarize performance over time. Benefits would include being able to identify performance trends, including recurring problems. The SMEs also felt that the tool could provide support for briefings up and down the chain of command, making it useful in a significantly wider variety of circumstances. Future work will include collecting additional feedback from representative users to insure that the continued development of the AAR is in line with O/C and commander needs.

CONCLUSIONS

The content and patterns of a team's communication provide a window into performance and cognitive states of the individuals and the team as a whole. By applying computational analyses of the communication stream, we can automatically derive team performance metrics. The feasibility of using this approach was demonstrated for automatically detecting critical incidents, identifying performance changes, and evaluating team performance in both live and virtual training environments.

The system uses a Statistical Natural Language-based intelligent software agent for embedding automatic, continuous, and cumulative analysis of spoken interactions in individual and team training and operational environments. Starting with an incoming stream of free-form verbal communication, commercial grade Automatic Speech Recognition (ASR) is applied, generating transcribed text and speech characteristics, such as voice stress, which can, in near real-time (within seconds), be analyzed using previously trained natural language models resulting in detailed measures of team characteristics and performance. This process provides a complete communications analysis pipeline, automatically converting team communications to performance metrics.

The DARCAAT toolkit allows the analysis and modeling of both objective and subjective performance metrics and is able to work with large amounts of communication data. The toolkit automatically extracts measures of performance by modeling how subject matter experts have rated similar communication as well as modeling objective performance measures. Because the technology uses automated machine-learning and natural language approaches, it does not require large amounts of hand-coded language analysis or task analysis. This permits rapid development of the technology for novel tasks and situations. Based on the success of this project, the AAR tool could be further developed into an operational tool for use in Ambush! and NTC STX lane training environments with some additional refinements.

ACKNOWLEDGEMENTS

We would like to acknowledge the contributions of the DARCAAT project team including Marita Franzke, Kyle Habermehl. Brent Halsey, Chuck Pannacione, Manju Putcha, Jim Parker, Boulder Labs, David Diller, Laura Leets, Fred Flynn, Cyle Fena, Don Scott, Paul Asuncion, Reginald King, Brian Oman, Len Dannhaus, Nick Hatchel, Rick Travis, David Leyden, and Jamison Winchel. We would also like to thank the following organizations for participating in this work: DARPA DSO, ARI, Fort Lewis, NTC, and Fort Carson. This work was sponsored by the Defense Adanced Research Pojects Agency (DARPA) and the U.S. Army Research Institute (ARI).

REFERENCES

Diller, D. E., Roberts, B., Blankenship, S. & Nielsen, D. (2004). DARWARS Ambush! – Authoring lessons learned in a training game. In Proceedings

of the Interservice/Industry Training, Simulation and Education Conference. Orlando, FL: I/ITSEC.

Diller, D. E., Roberts, B. & Willmuth, T. (2005). DARWARS Ambush! A case study in the adoption and evolution of a game-based convoy trainer with the U.S. Army. Presented at the Simulation Interoperability Standards Organization, 18-23 September.

Foltz, P. W. (2005). Tools for Enhancing Team Performance through Automated Modeling of the Content of Team Discourse. In Proceedings of HCI International, 2005.

Foltz, P. W., Laham, R. D. & Derr, M. (2003). Automated Speech Recognition for Modeling Team Performance. In Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.

Foltz, P. W., Lavoie, N., Oberbreckling, R. & Rosenstein, M. (2007). Tools for automated analysis of networked verbal communication. Network Science Report, Volume 1, 1 pp 19-24, United States Military Academy Network Science Center.

Foltz, P. W., Martin, M. A., Abdelali, A., Rosenstein, M. B. & Oberbreckling, R. J. (2006). Automated Team Discourse Modeling: Test of Performance and Generalization. In Proceedings of the 28th Annual Cognitive Science Conference.

Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. A. & Cooke, N. J. (2003). Evaluation of Latent Semantic Analysis-based measures of team communications content. In Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.

Hansen, J. H. L. et. al., (2002). Methods for Voice Stress Analysis and Classification, an appendix to Investigation and Evaluation of Voice Stress Analysis Technology, Final Report for National Institute of Justice, Interagency Agreement 98-LB-R-013. Washington, DC, 2002. NCJRS, NCJ 193832.

Jurafsky, J. & Martin, J. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, New York, Prentice Hall.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J & Martin, M. J. (2002). Some promising results of communication-based automatic measures of team cognition. Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting.

Kiekel, P., Gorman, J., & Cooke, N. (2004). Measuring Speech Flow of Co-located and Distributed Command and Control Teams During a Communication Channel Glitch. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, 683-687.

Kuhn, C. (2004). National Training Center: Force-on-force convoy STX lane. *Engineer: The professional bulletin for Army Engineers*, April-June.

Laham, D., Bennett, W., & Derr, M. (2002). Latent Semantic Analysis for career field analysis and information operations. Paper presented at Interservice/Industry, Simulation and Education Conference (I/ITSEC), December 2-5, 2002. Orlando, FL.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25(2&3), 259-284.

Lippold, O. (1971). Physiological Tremor, Scientific American, Volume 224 (3), March.

Schmidt-Nielsen, A., Marsh, E., Tardelli,, J., Gatewood, P., Kreamer, E., Tremain, T., Cieri, C., Strassel, S. Martey, N., Graff, D. & Tofan, C. (2001). Speech in Noisy Environments (SPINE2) Part 1 Audio. Linguistics Data Consortium catalog: LDC2001S04.

U.S. Department of the Army. (2001, June). FM 3-0: Operations. Washington, DC.

U.S. Department of the Army. (2003, September). FM 7-1: Battle focused training. Washington, DC.